Tecnologías de la Traducción: Mejoras a los sistemas por transferencia morfológica*

Felipe Sánchez Martínez, Mikel L. Forcada,
Carlos Pérez Sancho, Antonio Pertusa Ibáñez

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant

Introducción

Hay problemas de traducción que los sistemas por transferencia morfológica (también llamados sistemas por transferencia sintáctica parcial) que vimos en la primera parte de la práctica y que en clase denominamos *modelo 1* no son capaces de resolver con una cantidad de datos razonable. Este documento trata las limitaciones de este tipo de sistemas de TA basados en reglas o conocimiento lingüístico y hace un esbozo de las posibles estrategias de solución.

Actividad

Indicad algunos problemas no resueltos por el *modelo 1*, propuesto en la primera parte de la práctica, y haced un esbozo de las posibles estrategias de solución que se tendrían que incluir en un modelo más avanzado. Para inspiraros, fijaos en las siguientes grupos de frases y las traducciones producidas por sistemas por transferencia morfológica (*modelo 1*).
Grupo A: (traducciones de Apertium)

1. (ca) *Aquests avions no volen bé* → (es) **Estos aviones no quieren bien*
2. (es) *Que salen la sopa, que está sosa* → (ca) **Que ixen la sopa, que està insulsa*

Grupo B: (traducciones de Apertium)

1. (es) *Una almohada* → (ca) *Un coixí*
2. (es) *Una almohada cómoda* → (ca) *Un coixí còmode*
3. (es) *Una almohada más cómoda* → (ca) *Un coixí més còmode*
4. (es) *Una almohada bastante más cómoda* → (ca) **Un coixí bastant més còmoda*
5. (es) *Una almohada muy cómoda* → (ca) *Un coixí molt còmode*
6. (es) *Una almohada azul muy cómoda* → (ca) **Un coixí blau molt còmoda*
7. (es) *Una almohada de lana pero más cómoda* → (ca) **Un coixí de llana però més còmoda*

Grupo C: (traducciones de Apertium)

1. (en) *A house* → (ca) *Una casa*
2. (en) *A car* → (ca) *Un cotxe*
3. (en) *Red houses* → (ca) *Cases vermelles*
4. (en) *A large house* → (ca) *Una casa gran*
5. (en) *The young expert* → (ca) *L'expert jove*
6. (en) *The professor's house* → (ca) *La casa del professor*
7. (en) *The young professor's car* → (ca) *El cotxe del professor jove*
8. (en) *The professor's new house* → (ca) *La casa nova del professor*
9. (en) *The young professor's large car* → (ca) **El professor jove cotxe gran*

Notas para seguir la explicación del profesor

Descripción de los problemas y esbozo de la solución:

Grupo A

1. Como la palabra *volen* es una palabra homógrafa y las dos interpretaciones de la palabra son verbos (*volar*.verb.indic.pres.3.pl y *voler*.verb.indic.pres.3.pl) no hay manera de distinguirlos usando sólo información sobre las categorías que los acompañan. Una posible solución es refinar las categorías léxicas para distinguir los verbos intransitivos (*volar*) de los verbos transitivos (*voler*), o bien los verbos *léxicos* (*volar*) de los verbos *modales* (*voler*) y esperar que la distribución de las categorías de las palabras vecinas ayuden a distinguir un caso del otro. La otra consistiría en hacer un análisis más profundo que descartara uno de los dos casos.

2. Como en el ejemplo anterior, las dos interpretaciones de la palabra *salen* son verbos (*salar*.verb.subj.pres.3.pl y *salir*.verb.indic.pres.3.pl). También como en el ejemplo anterior, o bien tenemos que refinar las categorías léxicas y distinguir verbos en subjuntivo de los verbos en indicativo o los verbos transitivos de los intransitivos y esperar que las categorías léxicas de las palabras vecinas prioricen la interpretación correcta, o bien optar por un análisis más profundo que permitiera decidir en que caso nos encontramos.

En general, no es fácil resolver al nivel del *modelo 1* algunos tipos de **ambigüedad léxica**, como por ejemplo la homografía con coincidencia de categoría gramatical ((es) *fui*, *creo*, (ca) *podem*, *volem*) o la polisemia ((es) *destino* $\rightarrow$ (ca) *destí*, *destinació*), fijándonos sólo en las categorías de las palabras vecinas.

Grupo B

1. Si el sistema aplicara el *modelo cero* (un *modelo 1* sin reglas), el resultado habría sido *una coixí*. Una manera de explicar la traducción en el *modelo 1* sería decir que existe la regla:

$$
\begin{array}{c}
R_1\text{: (es) } \textbf{det subst} \rightarrow \text{(ca) } \textbf{det subst} \\
\text{asigna género meta: } \textbf{subst} \rightarrow \textbf{det} \\
\text{asigna número meta: } \textbf{subst} \rightarrow \textbf{det}
\end{array}
$$

2. Ahora podemos establecer la hipótesis de que el sistema es un *modelo cero* pero con la regla R_1; es decir, un sistema *modelo 1* con una regla sólo. Como la frase contiene el patrón que encontramos en la parte izquierda de la regla R_1, se aplicaría esta regla y el resto se traduciría con *el modelo cero*. El resultado habría sido *Un coixí cómoda* (la palabra *cómoda* se traduciría aisladamente, puesto que no la cubre la regla). Una manera de explicar la traducción observada en el *el modelo 1* sería decir que hay otra regla:

R_2: (es) **det subst adj** $\rightarrow$ (ca) **det subst adj**
asigna género meta: **subst** $\rightarrow$ **det**
asigna número meta: **subst** $\rightarrow$ **det**
asigna género meta: **subst** $\rightarrow$ **adj**
asigna número meta: **subst** $\rightarrow$ **adj**

3. A la vista de las observaciones anteriores, podemos establecer la hipótesis de que el sistema es un *modelo 1* que sólo tiene las reglas R_1 y R_2. Como la frase contiene el patrón que encontramos en la parte izquierda de la regla R_1 pero no el de la regla R_2, se aplicaría R_1 y el resto se traduciría con *el modelo cero*. El resultado habría sido *Un coixí més cómoda* (las palabras *més* y *cómoda* se traducirían aisladamente, puesto que no las cubre la regla). Una manera de explicar la traducción observada en el *el modelo 1* sería decir que hay una tercera regla:

   R_3: (es) **det subst adv adj** $\rightarrow$ (ca) **det subst adv adj**
   asigna género meta: **subst** $\rightarrow$ **det**
   asigna número meta: **subst** $\rightarrow$ **det**
   asigna género meta: **subst** $\rightarrow$ **adj**
   asigna número meta: **subst** $\rightarrow$ **adj**

   Es decir, la regla R_2 insertando el adverbio. Fijaos que el hecho de que las reglas estén basadas en secuencias de palabras y no en estructuras sintácticas hace que tengamos que hacer reglas específicas para cada posible secuencia, a pesar de las evidentes analogías.

4. Llegados a este punto podemos establecer la hipótesis de que el sistema es un *modelo 1* con tres reglas: R_1, R_2 y R_3. Como la frase contiene el patrón que encontramos en la parte izquierda de la regla R_1 pero no el de la regla R_2, se aplicaría R_1 y el resto se traduciría con *el modelo cero*, puesto que no se puede aplicar ninguna regla. El resultado es, como se observa, *Un coixí bastant més cómoda* (las palabras *bastant*, *més* y *cómoda* se traducirían aisladamente, puesto que no las cubre ninguna regla). Para producir la traducción más adecuada *Un coixí bastant més còmode* habría que introducir una cuarta regla:

   R_4: (es) **det subst adv adv adj** $\rightarrow$ (ca) **det subst adv adv adj**
   asigna género meta: **subst** $\rightarrow$ **det**
   asigna número meta: **subst** $\rightarrow$ **det**
   asigna género meta: **subst** $\rightarrow$ **adj**
   asigna número meta: **subst** $\rightarrow$ **adj**

Es decir, la regla R_3 insertando el segundo adverbio. Como antes, dado que las reglas son muy rudimentarias y están basadas en secuencias de palabras y no en estructuras sintácticas, tendríamos que hacer reglas específicas para cada posible secuencia.

5. El *modelo 1* con las tres reglas R_1, R_2 y R_3 explica perfectamente la traducción observada.

6. El *modelo 1* con las tres reglas R_1, R_2 y R_3 se queda corto puesto que aplicaría la regla R_2 a *Una almohada azul* y no encontraría reglas para aplicar a *muy cómoda*. Para tratar esta estructura haría falta una quinta regla:

R_5: (es) **det subst adj**$_1$ **adv adj**$_2$ $\rightarrow$ (ca) **det subst adj**$_1$ **adv adj**$_2$
asigna género meta: **subst** $\rightarrow$ **det**
asigna número meta: **subst** $\rightarrow$ **det**
asigna género meta: **subst** $\rightarrow$ **adj**$_1$
asigna número meta: **subst** $\rightarrow$ **adj**$_1$
asigna género meta: **subst** $\rightarrow$ **adj**$_2$
asigna número meta: **subst** $\rightarrow$ **adj**$_2$

7. Queda claro que en este caso sólo se puede aplicar la regla R_1 y no se puede asegurar la concordancia de la traducción del adjetivo *cómoda*, y que con reglas basadas en secuencias la única manera de asegurarla seria tener reglas que describan todas las posibles estructuras que se pueden encontrar entre *almohada* y *cómoda*, cosa que es claramente impracticable. Esta es claramente una limitación del *modelo 1*, el cual, en general, tiene problemas para tratar algunos casos de **concordancia**, como por ejemplo la concordancia "a la larga" ((es) *el postre que nos ofreció era delicioso* $\rightarrow$ (ca) **les postres que ens va oferir eren deliciós*, o la concordancia "a la corta" cuando el sintagma donde se tiene que establecer no corresponde a ninguna de las secuencias de palabras detectadas (cómo en el ejemplo de *almohada–coixí*).

Grupo C

1. El *modelo 1* con la regla R_1 explica perfectamente la traducción observada.

R_1: (es) **det subst** $\rightarrow$ (ca) **det subst**
asigna género meta: **subst** $\rightarrow$ **det**
asigna número meta: **subst** $\rightarrow$ **det**

2. Tanto el *modelo cero* como el *modelo 1* con la regla R_1 darían el mismo resultado, que es el observado: *un cotxe*.

3. Ahora podemos establecer la hipótesis de que el sistema es un *modelo cero* pero con la regla R_1; es decir, un sistema *modelo 1* con una regla sólo. Como la frase no contiene el patrón que encontramos en la parte izquierda de la única regla R_1, no se se aplicaría ninguna regla y la fase se traduciría con *el modelo cero*. El resultado habría sido *Vermell cases*. Pero la traducción observada es diferente. Una manera de explicar la traducción observada en el *modelo 1* sería decir que hay otra regla:

   R_2: (en) **adj subst** $\rightarrow$ (ca) **subst adj**
   asigna género meta: **subst** $\rightarrow$ **adj**
   asigna número meta: **subst** $\rightarrow$ **adj**

4. Llegados a este punto podemos establecer la hipótesis de que el sistema es un *modelo 1* con dos reglas: R_1 y R_2. Si asumimos que el sistema va de izquierda a derecha y en cada punto aplica la regla más larga que concuerda con la parte del texto que queda por procesar, y no aplica nunca más de una regla a la misma palabra (para evitar ambigüedades a la hora de aplicar las reglas), encontraremos que como no tiene ningún patrón que concuerde empezando en la primera palabra *A*, traducirá esta palabra aisladamente (*Un*). Después, la regla R_2 concuerda con el resto de la frase *large house* y la aplica. El resultado es *Un casa gran*. Para producir la traducción observada *Una casa gran* tenemos que postular la existencia de una nueva regla:

   R_3: (en) **det adj subst** $\rightarrow$ (ca) **det subst adj**
   asigna género meta: **subst** $\rightarrow$ **det**
   asigna número meta: **subst** $\rightarrow$ **det**
   asigna género meta: **subst** $\rightarrow$ **adj**
   asigna número meta: **subst** $\rightarrow$ **adj**

   Es decir, una especie de combinación de las reglas R_1 y R_2 en una sola regla.

5. El resultado de aplicar la regla R_3 deducida en la frase anterior sería idéntico al observado: no hay que postular ninguna regla nueva.

6. Si asumimos ahora que nuestro sistema es un *modelo 1* con las tres reglas R_1, R_2 y R_3, la única regla que concuerda con la frase (que podemos analizar **det subst gs subst**, donde **gs** representa la partícula del genitivo sajón *'s*, *s'* o ') es la R_1. El resto de la frase se traduciría con *el modelo cero* y el resultado sería *El professor casa* (donde suponemos que el sistema no asigna ninguna traducción a un genitivo sajón suelto). Pero este no es el resultado observado, para explicarlo tenemos que postular una nueva regla:

R_4: (en) **det**$_1$ **subst**$_1$ **gs** **subst**$_2$ $\rightarrow$ (ca) "el" **subst**$_2$ "de" **det**$_1$ **subst**$_1$
asigna género meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna número meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna género meta: **subst**$_2$ $\rightarrow$ "el"
asigna número meta: **subst**$_2$ $\rightarrow$ "el"

La forma "el" representa un artículo determinado y la forma "de" la preposición *de*. Fijémonos que se tienen que generar palabras nuevas para traducir la estructura.

7. Si asumimos ahora que nuestro sistema es un *modelo 1* con las cuatro reglas propuestas, R_1, R_2, R_3 y R_4, la única regla que concuerda con la frase (que podemos analizar **det adj subst gs subst**, donde **gs** representa la partícula de genitivo sajón *'s*, *s'* o *'*) es la R_3. Como en el ejemplo anterior, el resto de la frase se traduciría con el *modelo cero* y el resultado sería *El professor jove cotxe* (de nuevo, suponemos que el sistema no asigna ninguna traducción a un genitivo sajón suelto). Pero este no es el resultado observado, para explicarlo tenemos que postular una nueva regla, que es básicamente una variación de la regla R_4:

R_5: (en) **det**$_1$ **adj**$_1$ **subst**$_1$ **gs** **subst**$_2$ $\rightarrow$ (ca) "el" **subst**$_2$ "de" **det**$_1$ **subst**$_1$ **adj**$_1$
asigna género meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna número meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna género meta: **subst**$_1$ $\rightarrow$ **adj**$_1$
asigna número meta: **subst**$_1$ $\rightarrow$ **adj**$_1$
asigna género meta: **subst**$_2$ $\rightarrow$ "el"
asigna número meta: **subst**$_2$ $\rightarrow$ "el"

8. Ahora tenemos un sistema con cinco reglas, de la R_1 a la R_5. Las reglas más largas que de izquierda a derecha concuerdan con la frase (que podemos analizar **det subst gs adj subst**, donde **gs** representa la partícula del genitivo sajón *'s*, *s'* o *'*) son, primero, la R_1, que concuerda con **det subst**, y, después de saltarse el **gs**, la regla R_2, que concuerda con **adj subst**. Pero este no es el resultado observado, para explicarlo tenemos que postular una nueva regla, que es, de nuevo, una variación de la regla R_4:

R_6: (en) **det**$_1$ **subst**$_1$ **gs** **adj**$_2$ **subst**$_2$ $\rightarrow$ (ca) "el" **subst**$_2$ **adj**$_2$"de" **det**$_1$ **subst**$_1$
asigna género meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna número meta: **subst**$_1$ $\rightarrow$ **det**$_1$

asigna género meta: **subst**$_2$ $\rightarrow$ “el”
asigna número meta: **subst**$_2$ $\rightarrow$ “el”
asigna género meta: **subst**$_2$ $\rightarrow$ **adj**$_2$
asigna número meta: **subst**$_2$ $\rightarrow$ **adj**$_2$

Como se puede ver, según si aparecen adjetivos calificando el primer sustantivo de la construcción o el segundo, tenemos que escribir reglas nuevas que gestionan explícitamente estos adjetivos, a pesar de que queda claro que la operación general que representan las reglas R_4, R_5 y R_6 es, aproximadamente:

(en) **det**$_1$ GN$_1$ **gs** GN$_2$ $\rightarrow$ (ca) “el” GN$_2$ “de” **det**$_1$ GN$_1$

donde cada GN indica un *grupo nominal* formato por un **subst** y, opcionalmente, un **adj**. No se indica, pero habría que asegurar la concordancia dentro de cada uno de los grupos nominales y con los determinantes que los acompañan.

9. Finalmente, este es el ejemplo que faltaba de los descritos por la regla general que hemos comentado hace poco: cada uno de los dos sustantivos del grupo nominal va acompañado de un adjetivo. El sistema no la traduce bien (aplica la regla R_3 y después la R_2) porque faltaría una regla:

R_7: (en) **det**$_1$ **adj**$_1$ **subst**$_1$ **gs** **adj**$_2$ **subst**$_2$ $\rightarrow$ (ca) “el” **subst**$_2$ **adj**$_2$“de” **det**$_1$ **subst**$_1$ **adj**$_1$
asigna género meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna número meta: **subst**$_1$ $\rightarrow$ **det**$_1$
asigna género meta: **subst**$_2$ $\rightarrow$ **adj**$_1$
asigna número meta: **subst**$_2$ $\rightarrow$ **adj**$_1$
asigna género meta: **subst**$_2$ $\rightarrow$ “el”
asigna número meta: **subst**$_2$ $\rightarrow$ “el”
asigna género meta: **subst**$_2$ $\rightarrow$ **adj**$_2$
asigna número meta: **subst**$_2$ $\rightarrow$ **adj**$_2$

Como podemos ver, en el *modelo 1*, si no podemos identificar estructuras más generales (más abstractas) como los grupos nominales indicados más arriba, tenemos que escribir reglas que tratan explícitamente cada posible composición de la frase. Imaginemos ahora qué tendríamos que hacer con frases más complejas como *The [young] professor’s [black] student’s [large] car*. El orden de las palabras es inadecuado cuando el sintagma que se tiene que reordenar no corresponde a ninguna de las secuencias de palabras detectadas. Necesitaríamos estructuras todavía más abstractas, incluso jerárquicas (árboles de análisis sintáctico propiamente dichos). Tendríamos que ir a un modelo con más análisis, el *modelo 2*.

Necesidad del análisis sintáctico: Los problemas de los dos últimos grupos son debidos al hecho de que la concordancia y el orden de las palabras son fenómenos que no se explican bien a nivel de palabras sueltas o de secuencias de palabras y se tienen que explicar en términos de **sintagmas**, unidades de longitud variable y muchas veces de naturaleza recursiva (y, por lo tanto, de longitud indefinida) que las lenguas manipulan como una unidad. Se podría decir que la estructura de la lengua no es lineal, sino jerárquica y recursiva.

Como ya se ha sugerido, una posible aproximación a la resolución de estos problemas consistiría en hacer un análisis sintáctico posterior al morfológico, que permitiría determinar la estructura y la longitud de los sintagmas y manipular cada uno como una unidad. La fase de transferencia operaría ya no sobre secuencias de palabras sino sobre árboles de análisis sintáctico. El resultado es un sistema de traducción indirecta por *transferencia sintáctica*.

El análisis sintáctico necesita algoritmos muy especializados que ya no son tan rápidos como los de análisis morfológico. Además, es mucho más difícil escribir una especificación completa de la sintaxis de un idioma —que cubra prácticamente todas las frases posibles— que una especificación morfológica, y además, se hace explícito un nuevo problema: la ambigüedad *estructural* o *sintáctica*: una oración puede tener muchos árboles de análisis sintáctico, y hay que elegir uno.

Compromiso "inteligencia–memoria": No se debe olvidar que cada vez que un determinado nivel de análisis tiene problemas para abordar traducciones, hay dos alternativas:

- como se ha dicho hasta ahora, hacer un análisis más profundo (por ejemplo, pasar de análisis sintáctico parcial con patrones a un análisis sintáctico completo para tratar problemas de reordenamiento más complejos), o utilizar
- la fuerza *bruta*: muchos más datos (más reglas, reglas más especializadas para los casos concretos, etc.) sin aumentar el nivel de análisis (por ejemplo, añadir muchos más patrones de reordenamiento).

Debe haber un compromiso (que se da en otras muchas tareas) entre la *inteligencia* (complejidad) del modelo y la memoria (cantidad de datos) de que dispone.

Niveles de análisis: Estos son los *componentes* de las representaciones abstractas en función del nivel de análisis:

- *análisis morfológico:* categorías léxicas, información de flexión (género, número, persona, caso, tiempo, etc.)
- *análisis sintáctico:* constituyentes (sintagmas) o relaciones de dependencia que forman estructuras arbóreas, etc.

- *semántica*: roles semánticos (agente, paciente, beneficiario, etc.)